AFHRL-TP-82-35

(12)

AIR FORCE

HUMAN RESOURCES

ADA124167

COMPARISON OF LIVE AND SIMULATED ADAPTIVE TESTS

By

David R. Hunter

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

December 1982

Final Technical Paper

Approved for public release; distribution unlimited.

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

This paper has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

J. P. AMOR, Lt Col, USAF
Chief, Manpower and Personnel Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TP-82-35	2. GOVT ACCESSION NO. AD-A124 167	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* COMPARISON OF LIVE AND SIMULATED ADAPTIVE TESTS		5. TYPE OF REPORT & PERIOD COVERED Final
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR *(s)* David R. Hunter		8. CONTRACT OR GRANT NUMBER *(s)*
9. PERFORMING ORGANIZATION NAME AND ADDRESS Manpower and Personnel Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 77191810
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE December 1982
		13. NUMBER OF PAGES 58
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*		15. SECURITY CLASS *(of this report)* Unclassified
		15.a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of this abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

adaptive tests
computer adaptive testing (CAT)
computer simulation
simulation

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The purpose of this research was to compare test scores obtained from live administration of several adaptive tests with those scores obtained from the same sample of individuals through computer simulation of the adaptive tests. The use of computer simulation techniques for the evaluation of adaptive testing protocols has gained wide usage. However, the validity of these simulation techniques has not been established.

Three adaptive testing procedures were implemented, using two distinct item types. The adaptive testing procedures included (a) Two-Stage Adaptive Test, in which a 10-item routing test was followed by one of five 30-item

DD Form 1473 1 Jan 73 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

Item 20 (*Continued*)

measurement tests; (b) Pyramidal Adaptive Test, in which the subject was branched through a pyramidal structure of items until a total of 10 items were administered; and (c) Stratified Adaptive Test, in which items were selected from nine pools of items stratified on item difficulty. The two item types used were Word Knowledge and Visual Scanning. Approximately 400 subjects were tested on each type of item. In addition to the three adaptive tests, each subject was also tested using a 220-item test in a conventional format.

Item responses for each subject from the conventional test administration were used to generate simulated test scores for each of the three adaptive tests. These simulated scores were then compared with the scores obtained from the live adaptive test administrations, and both sets of scores were compared with scores from a 30-item conventional format test drawn from the 220-item test.

Results indicated that for both types of items, the simulated tests are not strictly parallel forms of the live tests. It was concluded that caution must be exercised in the use of computer simulation and that results from such procedures are not completely generalizable to live testing situations; however, the practical use of simulation was supported.

AFHRL Technical Paper 82-35 December 1982

COMPARISON OF LIVE AND SIMULATED ADAPTIVE TESTS

By

David R. Hunter

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

Reviewed and Submitted for Publication by

Lonnie D. Valentine, Jr.
Chief, Force Acquisition Branch

This publication is primarily a working paper.
It is published solely to document work performed.

PREFACE

The purpose of this research effort was to examine the use of computer simulation procedures for the evaluation of adaptive tests. The research is in support of the Force Acquisition and Distribution System subthrust, and Manpower and Force Management thrust. Additionally, this served as the dissertation of the author.

Master Sergeant Floyd Hudson, Airman John Garza, and Mr. Richard Nicewonger deserve special credit for their assistance in the collection of this data. The chairman, Dr. Benjamin Fruchter, and other members of the dissertation committee (Dr. Lonnie Valentine, Dr. John Loehlin, Dr. H. Paul Kelley, and Dr. E. Earl Jennings) are also due special thanks for their continued advice and support. Finally, for her untiring work in the many revisions of this manuscript, I am especially indebted to Mrs. Virginia Weems.

TABLE OF CONTENTS

LIST OF TABLES

COMPARISON OF LIVE AND SIMULATED ADAPTIVE TESTS

I. INTRODUCTION

Most tests of cognitive abilities employ a testing strategy in which all subjects receive the same set of items, subject to the constraints imposed by time limits. This set of items is usually comprised of items with a mean difficulty of .50 for the target population. This procedure results in a test that is efficient and discriminates well for subjects around the mean ability level but is considerably less discriminating and efficient for subjects beyond, say, one standard deviation from the mean ability level. This property of conventional tests has been described by Lord and Novick (1968) and forms the basis for the development of a new testing strategy which has variously been termed "adaptive" (Weiss & Betz, 1973), "branched" (Bayroff, 1964), "tailored" (Lord, 1970), "sequential" (Owen, 1969), and "response-contingent" (Wood, 1973). Under this strategy each individual being tested on a certain ability does not necessarily respond to the same set of items as do other individuals being tested. Rather, each individual responds to a set of items that has been selected so as to be appropriate for that person's particular ability level. In general, therefore, an individual of high ability would receive a set of items that, overall, is more difficult than the set of items received by an individual of lower ability.

One of the simplest examples of this strategy is the two-stage testing procedure in which all individuals take a common first-stage test. Then, on the basis of the scores achieved on the first-stage test, each individual is given one of several second-stage tests which differ in their overall difficulty. Obviously, the use of this type of testing strategy presents some difficulties in the assignment of scores to individuals--a person who gets 90 percent of the items in the most difficult test correct should certainly not receive the same score as a person who gets 90 percent of the items correct in the least difficult test.

Several procedures have been suggested for the scoring of adaptive tests. These include (a) percent correct, (b) difficulty of the last item attempted, (c) difficulty of the item that would have been presented after the last item, (d) average difficulty of all items attempted, and (e) average difficulty of all items answered correctly.

Additionally, several branching strategies have been investigated, including such procedures as the two-stage test described earlier, pyramidal tests in which an item pool structured like a pyramid is used, and mathematical procedures in which the item pool is unstructured. Generally, each of these procedures utilizes the individual's responses to items to select either more difficult or less difficult items for subsequent administration.

Often the scoring procedures and branching strategies, in various combinations, have been investigated through the use of computer simulations, using either Monte-Carlo-generated item responses or real item responses. This procedure has been favored for the evaluation of adaptive testing because of the prohibitively high costs associated with implementing adaptive testing procedures, typically requiring some sort of computer-based testing system. In a few studies, adaptive tests were actually implemented and response data collected under actual adaptive testing conditions (referred to hereafter as "live testing").

The results of computer simulation studies have been used extensively as the basis for evaluating adaptive procedures; however, there has been no attempt, thus far, to demonstrate the validity of these simulation techniques. The aim of this research was to investigate the validity of computer simulation techniques for the evaluation of adaptive testing procedures, through the comparison of adaptive test scores derived from real data computer simulations with test scores achieved by the same subjects under live testing conditions.

II. LITERATURE REVIEW

Background

A recently developed testing strategy, which has been closely linked with the development of modern high speed computational systems and low cost terminals, has been what is generally called "adaptive" testing or "tailored" testing (Green, 1970; Lord, 1970). In this testing paradigm, an attempt is made to match the test to the ability level of the subject. This is desirable because as Lord (1971a, 1971b, 1971d, 1980), among others (Green, 1970; Hick, 1951), has demonstrated, an ideal test (that is, one which maximizes the information gained from a given number of items) is one in which the subject answers 50 percent of the items correctly. In several theoretical studies (Lord, 1971a, 1971b, 1971d), the measurement effectiveness of an instrument has been shown to decline rapidly as the underlying ability level diverges from the mean ability level of the sample on which the test was normed.

By using an adaptive testing technique, the difficulty of the test can be matched more closely to the ability level of the subject. While this will not result in improved measurement for those subjects whose ability level is near the mean, Lord's theoretical studies have clearly demonstrated the improvement which may be expected in measurement of subjects whose ability levels are not close to the mean.

An early investigation of the use of adaptive testing in a comparison with conventional test procedures in a clinical setting was performed by Hutt (1947). He examined the relative effects upon IQ ratings of consecutive, as compared with adaptive methods of testing with the Revised Stanford-Binet. His results indicated that for the total population there were no significant differences between scores obtained by the two testing methods; however, for poorly adjusted individuals the adaptive methods yielded higher, and presumably more valid, IQ scores.

This finding may be a reflection of another advantage that has been suggested for adaptive testing--the exclusion of items that are much too difficult or much too easy for an individual. It has been suggested (Weiss & Betz, 1973) that individuals of low ability may become discouraged in conventional tests due to exposure to many items that are far beyond their level of ability. Since they may not put forth their full effort on those items which are appropriate for their ability level, these individuals may score even lower than might be expected.

Additionally, when guessing is possible, as it is in most multiple-choice format tests, the accuracy of the scores of the low ability person may be seriously decreased by wild guessing on the many items that are too difficult. Conversely, the scores of high ability subjects may be poorly estimated and subjected to additional variation due to the inclusion of many items that are far too easy. Boredom with

these items may decrease the level of effort of the high ability person, and the occurrence of "silly mistakes" on items that are far below the true ability level may also decrease the accuracy of the measurement.

These factors also may have been responsible for the results noted in a study of IQ assessment in an older population (aged 65 to 75) performed by Greenwood and Taylor (1965). They found that an adaptive administration of the Wechsler Adult Intelligence Scale (WAIS) resulted in a higher mean score than for a control group which took the WAIS under the conventional procedure.

Several approaches have been used in the study of adaptive testing. These might be broadly classified into empirical and simulation categories. The latter, perhaps because of the difficulties inherent in adaptive testing using paper-and-pencil techniques, has been the dominant mode of investigation of the properties and characteristics of adaptive testing.

Lord (1970, 1971a, 1971b, 1971d), in his theoretical studies of the measurement effectiveness of adaptive as compared to conventional testing, has made extensive use of simulated populations of persons with specified ability distributions. This approach has also been applied successfully by Waters and Bayroff (1971) and Weiss (1973, 1974), among others, in various evaluations of adaptive tests.

This simulation approach has the advantage of allowing the comparison of several modifications to the adaptive test on populations with varying ability distributions in a very short period of time. However, this approach cannot be used to assess adequately the interaction effects of real persons with an actual adaptive test, so that the conclusions must be taken as somewhat tentative until they can be evaluated by empirical studies.

Regrettably, to date there have been relatively few empirical studies of adaptive testing. In addition to the studies by Greenwood and Taylor (1965) and Hutt (1947), Wood (1969) has also investigated the use of paper-and-pencil adaptive tests. Wood developed adaptive tests of three lengths and administered them to 91 students. Using a criterion of course grades, he found correlations of about .35 for the three adaptive tests, which was not as great as the correlation between the criterion and a conventional test of considerably greater length. These negative findings may be confounded by the fact that the conventional test contained a heterogeneous set of items while the adaptive tests contained homogeneous items. Whether a homogeneous conventional test has greater validity is still an open question.

In an investigation of computer-based science testing, Hansen (1969) compared the validity and reliability of a 17-item adaptive test to a 20-item conventional test, both of which covered material found in a freshman physics course. In two experiments that used 56 and 30 students as subjects, the adaptive test was found to be

superior in both reliability and validity for prediction of final course grade and scores on an ability test. This was true even though in the adaptive test each student received only five items as compared to 20 items administered on the conventional test.

The saving in testing time which is possible by using adaptive testing is a major advantage in its favor, particularly when it yields equivalent reliability and validity. This was also noted by Ferguson (1969), who developed a model for computer-assisted adaptive testing and implemented and evaluated the model in an elementary school using a system of "Individually Prescribed Instruction." His model provided for adaptive testing of students for attainment of proficiency in arithmetic skills with presentation of new skill objectives or additional review of current skills being contingent upon identification of test outcomes. The results from this study showed that the computer-assisted model provided reliable results in substantially less time than did conventional methods.

Since the use of on-line computer-assisted instruction is a rapidly developing method of instruction (cf. Fletcher, 1975; Holtzman, 1970), the use of testing methods with minimal time requirements may prove very useful. The additional time which would have been consumed by conventional testing may be put to use in more productive ways (such as additional instruction or more in-depth diagnostic testing to identify specific difficulties).

Other investigators (Betz & Weiss, 1973; Larkin & Weiss, 1974) have also noted the advantage of adaptive tests over conventional methods in empirical comparisons. Studies simulating adaptive tests, but using real subject responses, have also produced comparable results.

In a series of studies (Cleary, Linn, & Rock, 1968a, 1968b; Linn, Rock & Cleary, 1969), response data from almost 5,000 subjects to 190 verbal items were used to evaluate conventional testing methods against a variety of adaptive tests. It was found that scores from these simulated adaptive tests when correlated against an external criterion (scores on the College Board Achievement Tests in American History and English Composition, and on the Verbal and Mathematical tests of the Preliminary Scholastic Aptitude Test) compared favorably with conventional tests of much greater length. In particular, it was estimated that one type of adaptive test could achieve, using about 17 items per subject, validity equal to a 190-item conventional test. However, considerable variation in the validity of the adaptive tests was found depending on the specific type of adaptive branching strategy and scoring procedure used.

Branching Strategies

Branching strategies refer to the rules by which items or groups of items are selected for administration to the subject. In a

conventional test, this strategy is simply that the person will start at the first item and proceed in a linear fashion until all items in the test are completed or the time limit for the test is reached. Furthermore, the sequence of items will be the same for all subjects.

In adaptive testing, however, more complex strategies may be utilized such that subjects may not receive the same items, or even the same number of items, but rather an attempt will be made to select from some item pool those items or sets of items which are most nearly appropriate for each individual being tested.

The simplest of these adaptive branching strategies, as previously mentioned, is the two-stage test. Using this strategy, a subject first takes a "routing" test, usually consisting of a relatively small number of items. Based on the score on this test, the subject is routed or branched to a "measurement" test which is appropriate for the subject's level of ability.

An empirical study performed by Angoff and Huddleston (1958) compared two-stage testing procedures with conventional tests on both verbal and mathematical abilities. Angoff and Huddleston used data on over 6,000 students who had taken the Scholastic Aptitude Test and scored the responses as though the students had taken first a routing and then a measurement test.

Angoff and Huddleston used a 40-item verbal routing test to determine which of two 36-item measurement tests should be administered. For mathematical ability, a 30-item routing test and two 17-item measurement tests were used. Results showed the measurement portions of the two-stage adaptive tests to be more reliable than the conventional tests and also to be more valid predictors of grade point averages.

In a theoretical study of two-stage testing, Lord (1971b) analyzed over 200 different two-stage strategies by examining the effects of different numbers of second stage measurement tests and varying lengths of routing and measurement tests. His best two-stage procedure consisted of an 11-item routing test which branched to one of six measurement tests, each containing 49 items. Basing his evaluation on an information function which he developed, Lord concluded that the two-stage test was as good as a 60-item test peaked around the mean of the ability distribution and provided increasingly better relative measurement as ability deviated from the mean. However, when guessing was assumed, the superiority of the adaptive tests declined, especially at the lower ability levels.

These findings were supported by a study by Betz and Weiss (1974) in which Monte Carlo simulation procedures were used to compare two-stage adaptive tests and a conventional ability test, and in an empirical study conducted by Larkin and Weiss (1975).

Multi-stage strategies have been developed based on a pyramidal or tree-like structure. In this model, individual items, or in some instances small groups of items with similar difficulties, are administered, and a branching decision is made depending on the subject's responses. In most instances, the first item to be administered would be an item of median difficulty. A correct response would then lead to the presentation of a more difficult item while an incorrect response would lead to the presentation of a less difficult item.

A variety of branching rules are possible with this model. The simplest would be an "up-one, down-one" rule in which the difficulty of the next item to be presented goes up one step for correct responses and down one step for incorrect responses; however, other rules which have been evaluated include "up-one, down-two," and "up-two, down-one" strategies.

When multiple items are used at each node of the pyramid, even more complex branching decisions which take into account the number of items answered correctly are possible. When items are of a type such that the alternatives can be ordered on correctness, then branching rules can be constructed which take into account the degree of error of the alternative selected.

Since part of the objective of the adaptive testing procedure is to arrive as quickly as possible at items which are appropriate for each examinee, the use of different entry points, using prior knowledge of the individual's approximate ability level, has been suggested as a means of eliminating the presentation of inappropriate items at the start of the test.

Another procedure which has been suggested is the use of large step intervals (the difference in difficulty between subsequent items) during the first portion of a test, followed by smaller step intervals later. This decreasing-step technique, which is sometimes referred to as the Robbins-Munro procedure, has been investigated by Lord (1971c) who showed, in theoretical analyses, that as the change in difficulty levels on subsequent items becomes smaller, better estimates of the underlying ability level are obtained. However, in a subsequent study, Lord (1971d) concluded that, while shrinking-step-size procedures have certain advantages, if more than six or seven items are to be administered to a subject, then the shortcuts required to keep the item pool of the shrinking method within reasonable bounds do not lead to better measurement than does the fixed-step method.

Weiss (1973) has developed an interesting procedure which he calls the stratified-adaptive test. Using this procedure, a large item pool is divided into several (Weiss uses nine) non-overlapping strata based on item difficulty. Examinees are routed from one stratum to another, using an up-one, down-one branching rule. A correct response leads to the selection of an item from the next more difficult stratum, while an incorrect response leads to the selection of an item from the next less difficult stratum. Within a stratum, the examinee is given the

most discriminating item not previously administered. Weiss suggests the use of differential entry points and variable length testing; however, other procedures may also be used.

The strategies described thus far can be described as having fixed-branching in that they use a structured item pool that has been constructed based principally on item difficulties. There also exist variable-branching adaptive models which, in contrast to the structured item pools of the fixed-branching models, require only item pools of known difficulties and discriminations.

According to Weiss and Betz (1973), in the variable-branching model,

> The general procedure consists of choosing each item in succession for each individual, based on his responses to all previous items, in order to maximize or minimize some measurement-dictated criterion for that individual. . . . Each item is selected by searching through the entire item pool of unadministered items to locate the next "best" item for that individual. (p. 36)

A Bayesian adaptive testing model which utilizes this approach has been developed by Novick (1969). Novick's model uses a regression-based approach which considers both information on the population of which the subject is a member and data acquired on the subject during the course of testing to determine item selection. Novick's contention that this Bayesian procedure would find its maximal usefulness for short tests has been supported by the work of Wood (Note 2), who used a different Bayesian procedure.

Non-Bayesian approaches to the use of variable-branching models have been evaluated, principally by Urry (1970) in a Monte Carlo study comparing adaptive tests with conventional tests.

Scoring Procedures

In an adaptive test different subjects may receive non-overlapping sets of items with considerably different mean difficulties; therefore, typical conventional test scoring procedures may be invalid. Consider, for example, a two-stage test in which one individual is routed to the high difficulty measurement test while another individual is routed to the low difficulty test, and in their respective measurement tests, they each answer 50 percent of the items correctly. It is intuitively obvious that the simple percentage correct score in this case is inappropriate, since the person who answered half of the more difficult questions correctly is not at the same ability level as the person who answered half of the easier questions correctly.

For this and related reasons, other scoring procedures have been developed which in most cases take into account the difficulty of the

items that constituted the particular test which an individual answered. Some of these scoring procedures may be used with any of the adaptive strategies described earlier, while others are applicable only to either a two-stage or a multi-stage test.

In two empirical studies of two-stage testing (Betz & Weiss, 1973; Larkin & Weiss, 1975), a scoring procedure was used which calculated maximum likelihood ability estimates for each individual based on a weighting of the difficulties of the items answered correctly in the routing and measurement tests. This method, which requires facilities for numerical operations that may not be available in a typical on-line instructional/testing system and which is more difficult to interpret than many other scoring methods, is based on the work of Lord (1970, 1971c), who has described the theoretical and mathematical bases of the procedure.

Another scoring procedure that is applicable to two-stage testing is simply to use the average difficulty of all items answered correctly. This procedure, which captures much of the essence of the more complicated maximum likelihood procedure, is relatively easy to compute and interpret.

The average difficulty score is also one of a number of scoring procedures that may be applied to multi-stage (pyramidal) tests. Other methods that have been suggested for scoring this type of adaptive test include:

1. Ordinal rank of the difficulty of the final item.

2. Terminal-Right-Wrong--which extends the ordinal rank score by taking into account the subject's performance on the last item.

3. Difficulty of the most difficult item answered correctly.

4. Difficulty of the final (*n*th) item.

5. Difficulty of the (*n* + 1)st item--which would take into account the subject's performance on the last (*n*th) item and would in effect add imaginary items to the pyramid.

6. Average difficulty of all items attempted (excluding the first item since it is attempted by all subjects).

7. All item scoring--a procedure developed by Hansen (1969) which assigns a score to all items in pyramidal tests, even those which the subject does not attempt, based on the subject's performance on the items presented. This procedure is based on the assumption that if an item of given difficulty is answered correctly, then all less difficult items would have also been answered correctly. Correspondingly, it is assumed that all items more difficult than an item answered incorrectly would also have been missed. Thus it is possible to assign a right/wrong score to all items in the test, based on only the few items actually administered.

A major deficiency of the present state of research is that there have been very few studies which compared these scoring procedures. Lord (1971d) made a theoretical comparison of the "final difficulty," "number right," and "average difficulty" scores and found that when the step size is fixed, the number-right score is perfectly correlated with the final difficulty score. He concluded that, "Although no optimum small-sample properties have been proven for the average difficulty score, it appears to be the score of choice for the up-and-down method at present" (Lord, 1971d, p. 10).

In the two principal empirical studies to date comparing different scoring procedures, Larkin and Weiss (1974, 1975) evaluated six different scoring procedures: (a) number correct, (b) mean difficulty--attempted, (c) mean difficulty--correct, (d) difficulty of final item, (e) difficulty of (n + 1)st item, and (f) all-item score. Their evaluation, which was based on 15-item pyramidal tests using only the "up-one, down-one" branching rule and constant step size, indicated that there were fairly high correlations between scores obtained from the various scoring procedures, but that the most stable scores were those obtained from the mean difficulty of all items attempted procedure, and the all-item scoring procedure. They concluded by saying that, "Pyramidal testing can provide estimates of ability which have stabilities comparable to those of longer conventional tests and greater than tests of the same length" (Larkin & Weiss, 1974, p. 43).

III. RESEARCH SPECIFICATIONS

Statement of the Problem

As noted in the review of the literature, many adaptive testing procedures have been proposed and, in a number of instances, these procedures have been compared, both with other adaptive procedures and with conventional tests, on the basis of data obtained from computer simulation studies. However, the validity of these simulation procedures has not been established and, indeed, has seldom been questioned. The assumption that data from computer simulations give rise to valid conclusions regarding adaptive testing procedures, therefore, needs to be tested, and its tenability resolved.

The purpose of the present study was to investigate the relationships between test scores obtained from computer simulations using real item responses and test scores obtained from administration of items in an adaptive testing format. To accomplish that goal, two large groups of examinees were tested using two item types--word knowledge and visual scanning (see Appendixes A and B for examples). Word knowledge items were chosen because much of the previous research involving adaptive testing has used this item type; thus it provides a link to the existing body of data. Visual scanning items were chosen because no previous work in adaptive testing has used this item type. The combination of the two item types, therefore, links this study to previous research and also extends the research in adaptive testing to a new type item. Considerations for the selection of item types suitable for adaptive testing will be presented in Section IV.

Examinees were tested using a long conventional test and three adaptive tests. Each examinee's item responses from the conventional test were used as input to a computer program which generated an estimate of the examinee's score on each of the three adaptive tests. Thus, each individual had scores from the three adaptive tests based on real responses and scores from the three adaptive tests based upon computer simulation. Comparisons of the scores obtained from the simulated adaptive tests and from the real response adaptive tests were then possible.

Experimental Hypotheses

In light of the previous studies, the following sets of general hypotheses were advanced to test the degree to which scores from live and simulated adaptive tests coincided and to evaluate the degree to which conclusions based on live adaptive tests paralleled those based on simulated adaptive tests.

Hypothesis 1: Parallel forms. Live adaptive tests and simulated adaptive tests are strictly parallel forms.

Hypothesis 2: Equivalent comparisons. Evaluations of the efficacy of an adaptive testing protocol are equally valid for both live and simulated sources of data.

Operational Definitions

Computer Simulation. The process by which a score or set of scores is generated for an individual by means of a computer program which, through reference to a set of item responses obtained from that individual, produces responses similar to those that the individual would have made to those same items had they been presented during a specified testing sequence.

Conventional Test. A test in which items are presented in a linear fashion, such that all examinees receive the same set of items in the same order.

Adaptive Test. A test in which the selection of items is tailored to the ability level of the individual being tested. Thus, given a pool of available items, not all examinees may receive the same set of items.

Two-Stage Adaptive Test. A test in which an individual's performance on a preliminary, routing test determines the selection of one of a number (greater than one) of possible subsequent measurement tests. In general, an individual who performs well on the routing test will receive a measurement test consisting of items of greater than average difficulty.

Pyramidal Adaptive Test. A test in which the pool of available items is arrayed in a pyramidal structure. In general, an individual begins with an item of median difficulty and is routed through the item pool to items of either greater or lesser difficulty based up his or her responses. That is, a correctly answered item will lead to an item of greater difficulty, while an incorrectly answered item will lead to an item of lesser difficulty.

Stratified Adaptive Test. A test in which the pool of available items is arranged into a number of strata based upon the item difficulty indexes, with no overlap between strata. An individual typically begins this test with the most discriminating item within the stratum of median difficulty and, based upon his or her response to that item, is administered the most discriminating item not previously administered in either the next more difficult stratum or the next less difficult stratum.

Strictly Parallel Forms. Two test forms are strictly parallel when their means, variances, covariances, and correlations with an outside criterion are not significantly different (Votaw, 1948).

Assumptions and Limitations

Listed below are the specific underlying assumptions for both the adaptive testing procedures and the statistical treatment of those measures used in this research.

Equivalent norming groups. The six groups of approximately 1,000 subjects each (see Method section) used to generate the item difficulty parameters for the word knowledge items are assumed to be equivalent.

Mode of presentation. Item difficulty parameters obtained for the word knowledge items using a paper-and-pencil test format are assumed to be identical to, or possibly a linear transformation of, the difficulty parameters of the same items presented via a computer terminal.

Unidimensionality. Each item type (word knowledge and visual scanning) is assumed to measure a single, unidimensional trait.

Local independence. An individual's response to any given item is independent of that person's response to any other given item.

Statistical assumptions. The typical assumptions pertaining to correlation were made. That is, the relationship between the two variables under consideration was regarded as approximately rectilinear, and pairs of observations on any one subject were assumed to be independent of pairs of observations for all other subjects (Guilford & Fruchter, 1978).

Generalization. The results of this study are limited to the population from which the sample was drawn and to the specific combinations of adaptive testing protocols and item types employed.

IV. METHOD

Subjects

Air Force basic trainees were selected as the target population for this study. A sample of approximately 12,000 enlisted personnel attending the Basic Military Training School at Lackland AFB were used in the generation of item parameter norms for the word knowledge items used in this study. An additional 844 basic trainees were used in the experimental testing. Of that number, 409 received the word knowledge tests, and 435 received the visual scanning tests. All testing was accomplished while the subjects were detailed to the Manpower and Personnel Division of the Air Force Human Resources Laboratory for approximately 4 hours during their 6th day of training.

Approximately 20 subjects per day (10 in the morning and 10 in the afternoon) were randomly drawn from the 160 to 200 basic trainees made available each day for experimental testing and assigned to this study. The sample had a median age of 18 years and was 73 percent male and 27 percent female.

Equipment

All tests were administered by computer. Ten identical testing stations were used for test administration. Each station consisted of one cathode-ray tube display (Model VR-17C), a typewriter-like keyboard (Model LK-40), two joysticks, and a special function keyboard. The joysticks and special function keyboard were not used in this study and the subjects were instructed not to use them. Each station was controlled by a minicomputer (Model PDP-11/04). The 10 minicomputers controlling the test stations were in turn connected to a host computer system (Model PDP-11/34) which provided a mass-storage capability and controlled the loading and execution of programs in the test stations' minicomputers. Data collected at the test stations were transferred to the mass-storage device of the host computer and later transcribed to magnetic tape for transmittal to the computer system used for data analysis.

All the equipment used for the administration of the test procedures and collection of data were manufactured by Digital Equipment Corporation, Maynard, Massachusetts. The operating system software and host-to-satellite communications software (RT-11 and Remote-11, respectively), were also developed by Digital Equipment Corporation.

Experimental Test Development

Two distinct domains of ability were chosen for use in this study. These domains were verbal ability as assessed by word knowledge items (Appendix A) and perceptual speed as assessed by a visual scanning task (Appendix B). Two types of items were chosen so as to expand the generalizability of the results of this study. The choice of the word knowledge item type was dictated by the predominance of this item type

in live testing implementations of adaptive testing procedures. Thus, there is a substantial body of literature dealing with word knowledge adaptive tests, and the inclusion of this item type allows for a direct comparison between these and previous results (cf. McBride & Weiss, 1974).

The choice of a visual scanning task as the second item type to be used was directed by several considerations. In order to improve generalizability, it was desirable to have as divergent an item type as possible from the word knowledge items. Additionally, the development and implementation of a new item type in an adaptive format would broaden the base of research using adaptive tests. Certain operational considerations were also taken into account in the selection of the second item type. These considerations were (a) the requirement for the existence of an item pool in excess of 500 distinct items; (b) on the average, each item could require no more than 30 seconds so that the needed number of items could be administered within the available time limits; and (c) the items should be essentially unifactorial.

Two additional types of items met these criteria and were evaluated before the visual scanning items were selected. Both rotated figures and digit span tests were constructed and found to be unsuitable for administration in an adaptive format. The rotated figures items lacked variability in the item difficulty parameter and hence were unsuitable for use in a process which relies on the availability of items with a large range in difficulty. The digit span test also proved to be lacking in item difficulty variability. Preliminary studies using the visual scanning items, however, demonstrated that item difficulty could be reliably controlled through manipulation of the display time allowed for search and the size of the array within which the targets were contained.

Three adaptive testing strategies were chosen for implementation: Two-Stage Adaptive Test, Pyramidal Adaptive Test, and Stratified Adaptive Test. These particular strategies were chosen principally because of the extent of their use in previous research. Tables 1, 2, and 3 list the studies which have dealt with Two-Stage, Pyramidal, and Stratified Adaptive tests, respectively, while Table 4 lists studies which have dealt with all other types of adaptive tests. In those tables, studies which simulated the adaptive tests by using item response data obtained from conventional tests are listed under the heading of Simulation, while studies which used entirely synthetic data are listed under the heading of Monte Carlo.

It may be seen from those tables that a great deal of research has been conducted using those three testing strategies. Indeed, more studies have used the Stratified Adaptive Test than any other single strategy.

Table 1. Studies of Two-Stage Adaptive Tests

Live Data Studies	Simulation Studies	Monte Carlo Studies
Betz & Weiss, 1973	Cleary, Linn, & Rock, 1968a	Betz & Weiss, 1974
Larkin & Weiss, 1975	Cleary, Linn, & Rock, 1968b	
	Linn, Rock, & Cleary, 1969	

Table 2. Studies of Pyramidal Adaptive Tests

Live Data Studies	Simulation Studies	Monte Carlo Studies
Bayroff & Seeley, 1967	Linn, Rock, & Cleary, 1969	None Reported
Larkin & Weiss, 1974		
Larkin & Weiss, 1975		
Hornke & Sauter, 1979		

Table 3. Studies of Stratified Adaptive Tests

Live Data Studies	Simulation Studies	Monte Carlo Studies
Weiss, 1973	None Reported	Vale & Weiss, 1975a
Waters, 1975a		
Waters, 1975b		
Vale & Weiss, 1975b		
Betz & Weiss, 1976		
Pine, 1977		
Bejar, 1977		
Sapinkopf, 1977		
Betz, 1977		
Waters, 1977		
Bejar, Weiss, & Gialluca, 1977		
Prestwood, 1979		
Kingsbury & Weiss, 1979		
Thompson & Weiss, 1980		
Gialluca & Weiss, 1980		

Table 4. Studies of Other Adaptive Tests

Live Data Studies	Simulation Studies	Monte Carlo Studies
Betz & Weiss, 1975	Jensema, 1974	Jensema, 1974
Cliff, Cudeck, & McCormick, 1977	Cliff, Cudeck, & McCormick, 1977	McBride, 1975a
Hansen, Ross, & Harris, 1978a	Kalisch, 1979	Lord, 1975
Hansen, Ross, & Harris, 1978b	Kalisch, 1980	Betz & Weiss, 1975
Schmidt, Urry, & Gugel, 1978		Ireland, 1976
McBride, 1979		Jensema, 1977
Johnson & Weiss, 1979		McBride, 1977
Thompson & Weiss, 1980		Urry, 1977a
Sympson, Weiss, & Ree, 1981		Ree, 1977
		English, Reckase, & Patience, 1977
		Cliff, Cudeck, & McCormick, 1977
		Maurelli, (Note 1)
		Cudeck, McCormick, & Cliff, 1979
		Kalisch, 1979
		Kalisch, 1980
		Ree, 1981

For each of the adaptive testing strategies (both live and simulated), the score produced for each individual was the average difficulty index of all items answered correctly. This scoring procedure has been used extensively (cf. Lord, 1971b; Larkin & Weiss, 1974; McBride, 1975b) and allows for direct comparability among all three adaptive tests. The score used for the conventional tests was the percentage correct.

Word Knowledge Items. From the pool of items maintained by the Air Force Human Resources Laboratory for the generation of experimental tests, 500 word knowledge items selected had an approximately rectangular distribution of item difficulties and positive discrimination indexes. Each word knowledge item consisted of a stem word and five alternatives, one of which was a synonym for the stem. From among these 500 items, 20 items were selected and designated as anchor items to be used to link together the tests. These anchor items were all highly discriminating and had equally spaced difficulty levels. The remaining 480 items were divided into six sets of 80 items each so as to provide nearly equal distributions of difficulty and discrimination in each set. Six test booklets were then prepared, each booklet consisting of the 20 anchor items and one of the six sets of 80 unique items.

Each test booklet was administered, without time limit, to a sample of at least 2,000 basic trainees (mixed male and female). Data were collected over a 6-month period. No examinee took more than one booklet, and the typical completion time was less than 45 minutes.

The examinees' responses were recorded on computer-scannable answer sheets which were later submitted to a standard item analysis procedure. Table 5 gives the difficulty indexes for each of the 20 anchor items in each of the six test booklets. The high degree of correspondence among the six sets of difficulty indexes supported the assumption of equivalent groups; therefore, as suggested by Vale, Maurelli, Gialluca, Weiss, & Ree (1981), no further item linking procedures were attempted. Defective items (that is, those having a negative discrimination index) and duplicates were discarded resulting in a usable pool of approximately 440 items.

In order to arrive at equivalent item pools to be used in the live adaptive testing and the simulated testing, the 440 remaining items were sorted into 10 non-overlapping groups based on item difficulty, and then within each group were sorted again (in descending order) on the item discrimination index. Even numbered items within even-numbered groups were assigned to the pool to be used by the live adaptive testing, and the odd-numbered items within those groups were

Table 5. Difficulty Indexes* for Anchor Items

	Booklet					
Item No.	I	II	III	IV	V	VI
3	96	98	97	97	96	97
5	59	59	57	56	61	56
12	86	86	86	86	87	85
15	81	84	81	81	79	79
16	58	59	54	53	58	43
18	49	51	46	44	48	45
23	89	89	88	88	89	89
33	90	89	89	**	89	88
34	73	72	73	72	70	69
38	43	45	45	43	40	42
42	84	79	81	82	80	78
43	78	80	78	77	76	78
49	29	27	25	27	26	27
53	83	84	83	83	80	83
56	50	51	46	48	50	46
63	87	88	86	86	84	86
66	44	48	39	42	43	47
69	24	28	27	26	22	24
73	81	82	79	79	80	81
78	30	27	27	26	31	29

*Decimal points omitted
**Misprinted, correct response omitted.

assigned to the simulation pool, while the assignment process was reversed for the odd numbered groups. This ensured that each pool had an almost exactly parallel distribution of difficulty and discrimination.

Conventional Test. Items belonging to the simulation pool were sorted into 10 groups based upon item difficulty (group 1 comprising items with difficulties ranging from .99 to .90, etc.). A conventional format test was then constructed by taking successive items from each group so as to form a modified spiralling test in which each successive set of 10 items contained items ranging from very easy to very difficult. This process was employed so as to spread fatigue effects over items from all the difficulty ranges. The items belonging to this pool were then stored, in order, in a designated computer file. Appendix A contains a sample word knowledge item. The same format was used for both conventional and adaptive tests. From items comprising the conventional test 30 items were identified to form a short conventional test for use as a criterion in the evaluation of the adaptive testing procedures. None of the items in this test were used in the generation of simulated adaptive test scores.

Stratified Adaptive Test. Within the item pool to be used in the live testing, nine groups of items were formed by combining those items with difficulty indexes in the range .00 to .09 with those in the range .10 to .19, to form the nine strata to be used by the stratified adaptive test model. This process was chosen over resorting the entire live testing item pool into nine equal interval strata because of the limited number of items available in the high difficulty range.

Two-Stage Test. The 10 items to be used in the routing portion of the two-stage adaptive test were chosen so as to have an approximately rectangular distribution of difficulty and were among the most discriminating of those in the live-testing pool. Five measurement tests of 30 items each were formed by selecting the 30 most discriminating items in the ranges shown in Table 6, excluding those items used in the routing test.

Table 6. Difficulty Ranges of Two-Stage Measurement Tests and Associated Routing Test Score

Measurement Test	Difficulty Range	Routing Test Range*
1	.01 - .20	9 - 10
2	.21 - .40	7 - 8
3	.41 - .60	5 - 6
4	.61 - .80	3 - 4
5	.81 - .99	0 - 2

*Number of items answered correctly on the Routing Test. (An individual answering seven items correctly would be routed to Measurement Test 2.)

Pyramidal Test. The 55 items used in the pyramidal test were chosen so as to approximate as closely as possible an item structure with a median difficulty index of .50 and a step value of $(\pm)$ .05. The item difficulty structure obtained is shown in Table 7. Items at the top of any column were the most discriminating available.

Table 7. Live Pyramidal Test Item Difficulty Indexes*
(Word Knowledge)

									49									
								56		46								
							59		50		39							
						66		54		44		36						
					70		59		49		42		29					
				77		65		55		45		34		25				
			79		68		58		50		40		29		20			
		86		77		65		56		46		34		24		15		
	91		81		68		60		52		41		29		20		10	
95		85		77		64		54		46		35		25		15		06

*Decimal points omitted.

Visual scanning Items. Each visual scanning item consisted of an array of digits consisting of from one to five rows of five to 20 digits per row. Thus the smallest array (and the easiest item) consisted of one row containing five digits, while the largest array (and the most difficult item) consisted of five rows, each containing 20 digits. For each item the subject's task was to count the number of occurrences of a randomly selected digit. The target digit always occurred at least once in the array.

The sequence of events for a visual scanning item was (a) present the target digit on the CRT, and wait 5 seconds; (b) present the array, and wait 7 seconds; and (c) solicit the subject's response, waiting as long as it takes for the subject to complete the response. A sample item showing these three steps is given in Appendix B.

The relationship between array size, display time, and item difficulty was determined through a tryout process in which the array size was systematically varied as described in the previous paragraph, and the display exposure time was held constant. The object was to find a display time at which approximately 50 percent of the items would be answered correctly and a smooth linear relationship between array size and proportion correct would be obtained. Table 8 shows the percent of items correct at each increment in array size for display time of 7 seconds, based on a sample of 79 individuals (not included in the 844 experimental subjects) and five observations per individual at each increment in array size. Thus each obtained value is based upon 395 observations. The mean percent correct is 50.55, and the correlation between array size and percent correct is -.98 ($p < .05$).

Table 8. Calibration of Visual Scanning Items

Array Size	Percent Correct
5	94.4
10	95.4
15	94.4
20	94.2
25	85.3
30	80.2
35	76.7
40	71.1
45	61.5
50	44.6
55	41.0
60	28.3
65	44.8
70	20.0
75	22.5
80	18.2
85	15.4
90	11.4
95	12.1
100	6.3

Conventional Test. A 220-item conventional-format test was assembled which began with the smallest (easiest) array (one row of five digits) and incremented the array size by five digits per item until reaching the largest (most difficult) array (five rows of 20 digits), and then restarting with the smallest array until a total of 220 items was reached.

Stratified Adaptive Test. Ten strata were defined consisting of those items with array sizes of 10, 20, 30, . . . 100 digits. Administration began with an item having an array of 50 digits. A total of 30 items was administered.

Two-Stage Test. The routing test consisted of 10 items having array sizes of 5, 15, 25, 35, . . . 95 digits. Administration began with the least difficult item and proceeded to the most difficult item. Based on the proportion of items answered correctly in the routing test, one of five measurement tests was chosen. The number of digits in the arrays comprising each of the five measurement tests is shown in Table 9. Each measurement test consisted of 30 items.

Table 9. Array Size for Two-Stage Measurement Test
(Visual Scanning)

Measurement Test	No. of Digits in Array	Routing Test Range
1	85 - 100	9 - 10
2	65 - 80	7 - 8
3	45 - 60	5 - 6
4	25 - 40	3 - 4
5	5 - 20	0 - 2

Pyramidal test. Table 10 shows the number of digits in the array at each point in the pyramidal test. Note that under conditions of equal item discrimination, the conduct of the pyramidal test is identical to a stratified adaptive test when the number of strata is equal to the number of items to be administered in the pyramidal test.

Table 10. Array Size for Elements of the Pyramidal Test
(Visual Scanning)

									50									
								45		55								
							40		50		60							
						35		45		55		65						
					30		40		50		60		70					
				25		35		45		55		65		75				
			20		30		40		50		60		70		80			
		15		25		35		45		55		65		75		85		
	10		20		30		40		50		60		70		80		90	
5		15		25		35		45		55		65		75		85		95

Software

Software systems for the interactive administration of adaptive and conventional tests have been developed by DeWitt and Weiss (1974) and Cudeck, Cliff, and Kehoe (1977) using the FORTRAN programming language and by McCormick and Cliff (1977) using the APL language. Because of the particular hardware configuration used in this study, however, none of these systems could be used without extensive modification. Therefore, the development of new computer software suitable for the PDP-11 computer systems was required.

Four computer programs were developed using the FORTRAN IV programming language. Two programs (TEST A and TEST B) were designed for execution on the PDP-11 minicomputers used for test administration and data collection using the word knowledge and visual scanning items, respectively. The programs were essentially parallel in design and function, the major

difference being that TEST A obtained items for presentation from several structured data files while TEST B generated the arrays of digits used as items by a pseudo-random process.

The word knowledge items to be used in each live adaptive test and the conventional test were contained in separate data files. The ordering of items within each file was structured so as to allow the computer program TEST A to access the correct item as specified by the adaptive testing algorithm. The program took advantage of the sequential file structure and used several counters to maintain a running record of the number of items administered, the difficulty strata of the last administered (in the case of the Stratified Adaptive Test), and the subject's response to select the next item for administration.

No input files were required for the program (TEST B) which used the visual scanning items, since the items were generated using a specified algorithm which specified the difficulty of the item as a function of the size of the array of digits to be scanned, while keeping the time available for scanning constant.

The data files produced by TEST A and TEST B were used as input to SIM A and SIM B, respectively. Those programs were designed for execution on a UNIVAC 1108 computer system and generated the simulated adaptive test scores. SIM A and SIM B closely parallel their corresponding live testing programs, TEST A and TEST B, respectively. For the word knowledge simulation, several files were used which specified the structured item pools for each adaptive test to be simulated. After an item was selected from a pool, the file containing the subject's responses was searched until his or her response to that particular item was located. A similar process was used for the generation of the simulated visual scanning adaptive tests.

All computer programs were verified through hand scoring and tracing of the item selection algorithms for selected cases from both the word knowledge and visual scanning samples.

Procedure

In order to counterbalance for fatigue effects, the order of administration of the conventional and adaptive tests, for both the word knowledge and visual search items, was alternated. On even-numbered days, the conventional test was administered first, followed by the adaptive tests. On odd-numbered days, the process was reversed.

Two rest periods were provided during the testing. One rest period of 5 minutes was provided between the end of the conventional test administration (220 items) and the start of the adaptive tests, or just prior to the conventional test administration for those cases in which the adaptive tests were administered first. In addition, a 5-minute rest period was also provided at the midpoint of the

conventional test administration. Since the testing time for all procedures totals just over 2 hours, this resulted in a rest period occurring approximately every 40 minutes.

The data obtained for each subject during the test administrations are shown in Table 11. In addition to the summary data obtained for each test, a code number specifying the item administered, keyed alternative, subject's response, and the response latency was recorded for the word knowledge tests. For the visual scanning tests, the size of the array, target digit, number of occurrences of the target, and subject's response were recorded.

Table 11. Experimental Test Data

Title		Definition
Live Adaptive Tests		
L-1	Two-Stage Routing Score	Percentage correct
L-2	Two-Stage Measurement Score	Average Difficulty of Correct Items
L-3	Pyramidal Score	Average Difficulty of Correct Items
L-4	Stratified Adaptive Score	Average Difficulty of Correct Items
Simulated Adaptive Tests		
S-1	Two-Stage Routing Score	Percentage correct
S-2	Two-Stage Measurement Score	Average Difficulty of Correct Items
S-3	Pyramidal Score	Average Difficulty of Correct Items
S-4	Stratified Adaptive Score	Average Difficulty of Correct Items
Conventional Tests		
C-220	220-Item Conventional Score	Percentage correct
C-30	30-Item Conventional Score	Percentage correct

Simulation Score Generation

Word Knowledge Tests. From among the 220 items comprising the conventional test, items were selected that paralleled as closely as possible the difficulty and discrimination indexes of the items used in each live adaptive test. Thus, for the simulated routing test of the Two-Stage Adaptive Test, 10 items were selected having an approximately rectangular distribution of difficulty and the highest discrimination indexes. (Table 12 shows the correspondence between the live and simulated test item difficulty indexes for the routing test.)

Table 12. Live and Simulated Two-Stage Routing Test Item Difficulty Indexes* (Word Knowledge)

Live Test	Simulated Test
91	92
81	81
77	76
68	67
59	58
56	54
47	46
37	39
29	29
14	15

*Decimal points omitted.

Each subject's file of responses to the 220 items of the conventional test was examined to determine the subject's responses to the 10 items comprising the routing test. The number of correct responses from among that set of items was tallied and converted to a percentage of correct responses (Score = [Number Right/10] x 100) which became the simulated Two-Stage Routing Test score.

Based on the simulated Two-Stage Routing Test score, one of the five sets of items comprising the simulated measurement tests was chosen according to the rules given in Table 6. (Table 13 shows the correspondence between the live and simulated test item difficulty indexes for the five measurement tests.) The subject's responses to those items were then determined, and a simulated Two-Stage Measurement Test score was generated by computing the average difficulty index for all items answered correctly.

Table 13. Live and Simulated Two-Stage Measurement Tests Item Difficulty Indexes* (Word Knowledge)

Measurement Test									
1		2		3		4		5	
Live	Sim	Live	Sim	Live	Sim	Live	Sim	Live	Sim
06	09	22	22	41	41	61	61	81	81
09	09	23	25	41	41	61	62	81	81
10	10	23	25	42	44	62	63	82	82
11	11	24	26	43	45	64	63	84	83
14	12	25	26	44	46	64	63	84	83
15	12	25	27	44	46	65	63	85	84
15	12	26	28	45	47	65	64	85	84
15	12	27	28	46	49	65	65	86	84
16	13	27	29	46	49	65	65	86	85
16	15	28	29	46	49	66	65	86	85
17	15	29	29	47	49	66	66	87	85
17	16	29	30	47	50	68	66	87	86
17	16	29	30	47	50	68	66	87	88
18	16	29	30	49	50	69	67	88	88
18	16	30	30	49	50	70	67	89	88
18	17	31	31	52	51	72	67	89	88
19	18	32	32	53	52	72	69	90	89
19	18	33	32	53	53	72	71	91	89
19	18	34	33	53	54	73	71	92	90
19	18	34	34	54	55	74	72	92	91
19	18	36	34	54	55	75	72	93	91
20	18	37	35	55	56	76	73	93	92
20	19	37	36	56	56	76	73	93	92
20	20	37	36	56	57	77	74	94	92
20	21	37	36	56	57	77	76	94	92
21	22	37	37	58	57	78	76	95	93
22	24	38	37	58	58	78	78	95	94
24	24	39	38	59	58	79	79	96	95
24	25	40	40	60	60	79	79	96	96
26	26	40	40	60	60	79	79	98	96

*Decimal points omitted.

The simulated Pyramidal Test was constructed in a similar fashion. Items were identified in the conventional test that closely approximated the parameters of the items used in the live Pyramidal Test. (Table 14 shows the item difficulty indexes for items at each point in the simulated pyramidal test.) The adaptive testing logic described earlier for the Pyramidal Test was used to step through the subject's responses to each

item and select the next item for simulated administration. As in the live testing, the score produced by this process was the average difficulty of those items answered correctly.

Table 14. Simulated Pyramidal Test Item Difficulty Indexes*
(Word Knowledge)

									50									
								54		46								
							60		50		39							
						65		56		46		36						
					72		59		50		40		30					
				74		66		54		44		34		26				
			81		71		61		49		41		29		20			
		86		76		64		53		45		36		25		15		
	91		81		69		59		51		39		31		18		12	
95		85		76		65		55		46		35		26		12		09

*Decimal points omitted.

For the simulated Strati ied Adaptive Test, an item-for-item matching of the items contained in the live adaptive testing to those in the conventional test was not performed. Rather, the process which produced the strata used in the live Stratified Adaptive testing was reproduced using items from the conventional test. The 220 items from the conventional test were sorted into nine strata corresponding to the strata used in the live Stratified Adaptive testing, and within each strata were sorted into descending order based upon the item discrimination index. The Stratified Adaptive Test logic described earlier was then followed and branching decisions made based upon the subject's responses to the items administered in the conventional format. The score produced by this process was the average difficulty of those items answered correctly from among the 30 selected for simulated administration in the Stratified Adaptive process.

Visual Scanning Tests. The process followed in the generation of the simulated adaptive test scores for the visual scanning items was essentially identical to that described for the word knowledge tests. The principal exception lies in the assumption of equivalent item discrimination indexes for all visual scanning items of equal array size, which eliminated the necessity for any matching of items on other than the item difficulty parameter.

V. RESULTS

Parallel tests, according to Gulliksen (1950), have equal means, equal variances, equal intercorrelations, and equal validities for any given criterion. To address the first experimental hypothesis given in the Research Specifications section, corresponding live and simulated adaptive tests were compared on each of those parameters. In addition, overall comparisons of the live and simulated adaptive tests were performed. Votaw (1946) has described statistical criterion for parallel tests which simultaneously assesses the degree to which a set of tests has equal means, variances, covariances, and validities with some external criterion. Computation procedures for this statistic are given by Gulliksen (1950). For the case in point (two parallel tests and one criterion--performance on a 30-item linear test), the statistic is defined (Gulliksen, 1950, p. 185) as:

$$\hat{L}_{mvc} = \frac{s_y^2 s_1^2 s_2^2 (1 + 2r_{y1} r_{y2} r_{12} - r_{y1}^2 - r_{y2}^2 - r_{12}^2)}{(s_y^2(u + w) - 2\bar{c}_{yx}^2)(u - w + v)},$$

where $u = (s_1^2 + s_2^2) / 2,$

$w = r_{12} s_1 s_2,$

$v = (\bar{X}_1 - \bar{X}_2)^2 / 2,$

$\bar{c}_{yx} = (c_{y_1} + c_{y_2}) / 2.$

s^2 designates a variance,

$\bar{X}$ designates a mean of one form of Test X,

r designates a Pearson product-moment correlation coefficient, and

c designates a covariance.

Subscripts:
1 designates form 1 of Test X,
2 designates form 2 of Test X, and
y designates the criterion measure

The quantity $-N\log_e\hat{L}_{mvc}$ is reported in Table 15 for the three adaptive tests using word knowledge items and in Table 16 for the three adaptive tests using visual scanning items. When N is large and the null hypothesis (given in Tables 15 and 16) is true, then the quantity $-N\log_e\hat{L}_{mvc}$ is distributed approximately as chi-square with three degrees of freedom. Gulliksen (1950, p. 189) provides a table for the evaluation of this quantity at the 1 and 5 percent levels of significance. As shown in Tables 15 and 16, the null hypothesis of strictly parallel tests is rejected for all comparisons except the Pyramidal Adaptive Test using word knowledge items.

Table 15. Values of Votaw's Statistic $-N\log_e\hat{L}$ for the Three Tests of Compound Symmetry (Word Knowledge) ($\underline{N}$ = 409)

Test	Hypothesis $\hat{H}_{mvc}$	Hypothesis $\hat{H}_{vc}$	Hypothesis $\hat{H}_m$
Two-Stage Measurement Test	'3.5*	8.2*	***
Pyramidal Test	3.6	**	**
Stratified Adaptive Test	12.8*	7.7*	***

*$\underline{p}$ < .05
**Not computed because $\hat{H}_{mvc}$ failed to reach significance.
***Not computed because $\hat{H}_{vc}$ was rejected.

Hypothesis $\hat{H}_{mvc}$: Population means, variances, and covariances are equal, and population covariances with criterion are equal.

Hypothesis $\hat{H}_{vc}$: Population variances and covariances are equal and population covariances with criterion are equal.

Hypothesis $\hat{H}_m$: Population means are equal, given that $\hat{H}_{vc}$ is true.

Table 16. Values of Votaw's Statistic $-N\log_e\hat{L}$ for the Three Tests of Compound Symmetry (Visual Scanning) ($\underline{N}$ = 435)

Test	Hypothesis $\hat{H}_{mvc}$	Hypothesis $\hat{H}_{vc}$	Hypothesis $\hat{H}_m$
Two-Stage Measurement Test	11.7*	2.6	9.0*
Pyramidal Test	82.5*	2.3	80.2*
Stratified Adaptive Test	22.8*	14.0*	**

*$\underline{p}$ < .05
**Not computed because the $\hat{H}_{vc}$ was rejected.

Hypothesis $\hat{H}_{mvc}$: Population means, variances, and covariances are equal, and population covariances with criterion are equal.

Hypothesis $\hat{H}_{vc}$: Population variances and covariances are equal and population covariances with criterion are equal.

Hypothesis $\hat{H}_m$: Population means are equal, given that H_{vc} is true.

Having rejected the hypothesis of completely parallel tests, it is then possible to see if the differences in the variances and covariances of the parallel tests account for the failure to satisfy H_{mvc}. This statistic ($\hat{L}_{vc}$) is given in Gulliksen (1950, p. 187) as

$$\hat{L}_{vc} = \frac{s_y^2 \, s_1^2 \, s_2^2 \, (1 + 2r_{y_1} r_{y2} r_{12} - r_{y1}^2 - r_{y2}^2 - r_{12}^2)}{(s_y^2(u + w) - 2\bar{c}_{yx}^2)(u - w)},$$

where
$$u = (s_1^2 + s_2^2) / 2,$$
$$w = c_{12} = r_{12} s_1 s_2,$$
$$\bar{c}_{yx} = (c_{y_1} + c_{y_2}) / 2.$$

The quantity $-N \log_e \hat{L}_{vc}$ is reported in Tables 15 and 16 for the word knowledge and visual scanning tests respectively, for all but the Pyramidal Adaptive Test using word knowledge items. Comparison of the obtained values with the critical values provided by Gulliksen leads to the rejection of the null hypothesis ($\hat{H}_{vc}$) given in Tables 15 and 16 for the Two-Stage Adaptive Test and the Stratified Adaptive Test using word knowledge items, and for the Stratified Adaptive Test using visual scanning items.

For those instances in which the hypothesis $\hat{H}_{mvc}$ has been rejected, while the hypothesis H_{vc} has been sustained, it is then possible directly to test the notion that differences in the means of the parallel tests account for the failure to satisfy $\hat{H}_{mvc}$.

This statistic ($\hat{L}_m$) is given in Gulliksen (1950, p. 187) as

$$\hat{L}_m = \frac{(u - w)}{(u - w + v)}.$$

where the symbols are as previously defined.

The quantity $-N \log_e \hat{L}_m$ is distributed approximately as chi-square if N is large and $\hat{H}_m$ is true. However, direct interpretation

is possible only for those instances where $\hat{H}_{VC}$ is true. Therefore, rejection of the null hypothesis ($\hat{H}_m$) is possible only for the Two-Stage and Pyramidal Adaptive Tests using visual scanning items.

In addition to the use of Votaw's criterion, it is possible to compare the live and simulated adaptive tests directly using more traditional comparisons. Tables 17, 18, and 19 give the means, standard deviations, correlations with the criterion (30-item conventional test), and intercorrelations of the live and simulated Two-Stage Measurement Test, Pyramidal Test, and Stratified Adaptive Test, respectively, using word knowledge items. Tables 20, 21, and 22 present the same information for the live and simulated Adaptive tests using visual scanning items. The t-statistics reported were computed using the procedures for evaluating correlated means (Garrett, 1958, p. 226), correlated variances (Guilford & Fruchter, 1978, p. 170), and correlated correlations (Guilford & Fruchter, 1978, p. 167). These comparisons are in agreement with the results of analyses using Votaw's statistic.

Table 17. Comparison of Live and Simulated Two-Stage Measurement Tests (Word Knowledge) ($\underline{N}$ = 409)

	Live		Simulated	t
Mean	.4976		.4651	3.96*
Standard Deviation	.226		.210	2.11*
r_{C-30}**	.701		.643	3.19*
r_{LS}***		.712		

*$\underline{p}$ < .05
**Correlation with 30-item conventional test.
***Correlation between Live and Simulated tests.

Table 18. Comparison of Live and Simulated Pyramidal Tests (Word Knowledge) ($\underline{N}$ = 409)

	Live		Simulated	t
Mean	.5186		.5237	1.14
Standard Deviation	.109		.114	1.22
r_{C-30}	.692		.677	0.56
r_{LS}		.672		

Table 19. Comparison of Live and Simulated Stratified Adaptive Tests (Word Knowledge) (N = 409)

	Live	Simulated	t
Mean	.5491	.5408	2.29*
Standard Deviation	.163	.154	2.57*
$r_{C\text{-}30}$	.785	.773	0.88
r_{LS}	.895		

*p < .05

Table 20. Comparison of Live and Simulated Two-Stage Measurement Tests (Visual Scanning) (N = 435)

	Live	Simulated	t
Mean	.5731	.5980	2.93*
Standard Deviation	.1367	.1461	1.48
$r_{C\text{-}30}$	.254	.233	0.38
r_{LS}	.261		

*p < .05

Table 21. Comparison of Live and Simulated Pyramidal Tests (Visual Scanning) (N = 435)

	Live	Simulated	t
Mean	.5227	.5550	9.11*
Standard Deviation	.0635	.0598	1.32
$r_{C\text{-}30}$	.457	.475	0.39
r_{LS}	.325		

*p < .05

Table 22. Comparison of Live and Simulated Stratified Adaptive Tests (Visual Scanning) ($\underline{N}$ = 435)

	Live	Simulated	t
Mean	.5468	.5591	2.89*
Standard Deviation	.0873	.0764	3.12*
r_{C-30}	.559	.450	2.71*
r_{LS}	.453		

*$\underline{p}$ < .05

The preceding analyses have demonstrated that, except for the Pyramidal Adaptive test using word knowledge items, the live and simulated adaptive tests are not strictly parallel tests. Rather, there is a statistically significant effect of having a test administered in a live adaptive format that is reflected in differences in the mean level of performance and in some cases in the variance and correlation with a criterion.

While statistically significant effects were obtained, inspection of the live and simulated test comparisons (Tables 17 to 22) reveals that the magnitude of the effects is generally small and might not be sufficient to influence decisions made from using these data regarding the relative efficacy of differing adaptive tests. Therefore, to explore the second hypothesis given earlier, pairwise comparisons were performed contrasting the validity (correlation with 30-item conventional test) of each combination of adaptive test using both the live

Table 23. Comparison of Two-Stage Adaptive Test (Measurement Portion) with Pyramidal Adaptive Test

Item Type	Data Source	Test	r_{C-30}
Word Knowledge*	Live	Two-Stage	.701
		Pyramidal	.692
	Simulation	Two-Stage	.643
		Pyramidal	.677
Visual Scanning*	Live	Two-Stage	.254
		Pyramidal	.457**
	Simulation	Two-Stage	.233
		Pyramidal	.475**

*Live and Simulation comparisons agree.

**Superior Test ($\underline{p}$ < .05).

testing results and the simulated testing results. These comparisons are shown in Tables 23, 24, and 25, for the Two-Stage Adaptive Test versus Pyramidal Adaptive Test, Pyramidal Adaptive Test versus Stratified Adaptive Test, and Two-Stage Adaptive Test versus Stratified Adaptive Test, respectively, for both item types. Tables 26 and 27 give the complete intercorrelation matrices for the word knowledge and visual scanning tests, respectively, upon which these comparisons were based.

Table 24. Comparison of Pyramidal Adaptive Test and Stratified Adaptive Test

Item Type	Data Source	Test	r_{C-30}
Word Knowledge*	Live	Pyramidal	.692
		Stratified Adaptive	.785**
	Simulation	Pyramidal	.677
		Stratified Adaptive	.773**
Visual Scanning***	Live	Pyramidal	.457
		Stratified Adaptive	.559**
	Simulation	Pyramidal	.475
		Stratified Adaptive	.450

*Live and Simulation comparisons agree.
**Superior Test ($p < .05$).
***Live and Simulation comparisons disagree.

Table 25. Comparison of Two-Stage Adaptive Test (Measurement Portion) with Stratified Adaptive Test

Item Type	Data Source	Test	r_{C-30}
Word Knowledge*	Live	Two-Stage	.701
		Stratified Adaptive	.785**
	Simulation	Two-Stage	.643
		Stratified Adaptive	.773**
Visual Scanning*	Live	Two-Stage	.254
		Stratified Adaptive	.559**
	Simulation	Two-Stage	.233
		Stratified Adaptive	.450**

*Live and Simulation comparisons agree.
**Superior Test ($p < .05$).

Table 26. Correlations* Among Live and Simulated Adaptive Tests and Conventional Tests (Word Knowledge) ($\underline{N}$ = 409)**

	L-1	L-2	L-3	L-4	S-1	S-2	S-3	S-4	C-220	C-30
L-1	1.000									
L-2	.976	1.000								
L-3	.693	.681	1.000							
L-4	.846	.825	.774	1.000						
S-1	.745	.727	.675	.824	1.000					
S-2	.724	.712	.664	.798	.973	1.000				
S-3	.729	.715	.672	.799	.770	.742	1.000			
S-4	.837	.813	.756	.895	.863	.839	.869	1.000		
C-220	.849	.821	.792	.911	.819	.792	.833	.940	1.000	
C-30	.725	.701	.692	.785	.677	.643	.677	.773	.871	1.000

*All correlations reported are significant at $\underline{p} < .05$. Critical $\underline{r}$ = .098.

**See Table 11 for definition of symbols.

Table 27. Correlations* Among Live and Simulated Adaptive Tests and Conventional Tests (Visual Scanning) ($\underline{N}$ = 435)**

	L-1	L-2	L-3	L-4	S-1	S-2	S-3	S-4	C-220	C-30
L-1	1.000									
L-2	.928	1.000								
L-3	.240	.197	1.000							
L-4	.329	.302	.503	1.000						
S-1	.289	.261	.201	.272	1.000					
S-2	.283	.261	.181	.246	.945	1.000				
S-3	.277	.254	.325	.408	.288	.271	1.000			
S-4	.390	.364	.445	.453	.636	.600	.501	1.000		
C-220	.390	.358	.602	.671	.272	.246	.408	.453	1.000	
C-30	.277	.254	.457	.559	.272	.233	.475	.450	.770	1.000

*All correlations reported are significant at $\underline{p} < .05$. Critical $\underline{r}$ = .098.

**See Table 11 for definition of symbols.

From Tables 23 to 25 it may be seen that of the six pairs of comparisons, the evaluations based upon live adaptive tests were in agreement with the evaluations based upon simulated adaptive tests in five instances. The one instance in which agreement did not occur was in the comparison of Pyramidal and Stratified Adaptive Tests using visual scanning items. An evaluation based upon simulation data in that instance would have led to the conclusion that the Pyramidal and Stratified Adaptive tests were equivalent, while the live testing data show the Stratified Adaptive test to be superior, in terms of correlation with the 30-item conventional test.

VI. SUMMARY AND CONCLUSIONS

This study has addressed the adequacy of computer simulation techniques for the evaluation of adaptive testing procedures. The three adaptive tests chosen for implementation and evaluation have been studied extensively by other researchers and are typical adaptive testing techniques. The word knowledge items chosen represent a link to the extensive body of literature using this item type. The visual scanning items are a step toward greater use of the power of the computer for implementation of new, novel testing procedures such as those described by Church & Weiss (1980), Elwood & Griffin (1972), Hunter (1975, 1977), and Sanders, Valentine, & McGrevy (1971).

While no direct comparisons with other studies are available for the visual scanning test results, the results obtained from the word knowledge tests are comparable to those noted in the literature. Specifically, the correlations of the Two-Stage Adaptive Test, Pyramidal Adaptive Test, and Stratified Adaptive Test for both the live and simulated administrations with the 30-item criterion test were approximately in the .64-.70, .68-.69, and .77-.78 ranges, respectively. Previous studies dealing with live Two-Stage Adaptive Tests (Betz & Weiss, 1973), Pyramidal Adaptive Tests (Bayroff & Seeley, 1967; Hornke & Sauter, 1979; Larkin & Weiss, 1974), and Stratified Adaptive Tests (Thompson & Weiss, 1980; Vale & Weiss, 1975a; Waters, 1975a) have observed similar coefficients. The simulation studies of Two-Stage Adaptive Tests (Cleary et al., 1968a, 1968b; Linn, Rock, & Cleary, 1969) and Pyramidal Adaptive Tests (Linn et al., 1969) have generally reported somewhat higher correlations (.87-.95); however, these studies were based upon part/whole correlations in which the items contained in the adaptive tests were also contained in the criterion test. In addition, for the three studies cited, a 190-item criterion test was used. A Monte Carlo study reported by Betz & Weiss (1974) noted correlations in the range of .79-.82 between a 30-item Two-Stage Adaptive Test and a 40-item criterion test. The similarity of findings supports the generalizability of the obtained results.

Two hypotheses were explored. The first hypothesis addressed the question of whether the live adaptive tests and simulated adaptive tests were strictly parallel tests. That is, for each pair of tests, were there significant differences between their mean scores, variances, covariances, and correlations with an outside criterion? These comparisons were performed simultaneously by use of a statistic developed by Votaw (1948), and, in all cases, except for the Pyramidal Adaptive Test with word knowledge items, it was noted that the live and simulated tests were not strictly parallel. Individual comparisons revealed a significant difference between each pair of means. For the word knowledge Two-Stage and Pyramidal Adaptive Tests the live tests were slightly (but significantly) less difficult than their simulation counterparts. For all of the adaptive tests using the visual scanning items, live tests were slightly (but significantly) more difficult than their simulation counterparts. No explanation is available for the discrepancy between the two item types.

Less consistent differences were noted between the live and simulated tests' variances, covariances, and correlations with the outside criterion. Correlations between the live and simulated adaptive tests using the word knowledge items were consistently higher than the corresponding correlations based upon the visual scanning items. This may be attributed in part to the lower reliability of the visual scanning tests. The KR-20 index for the 30-item word knowledge criterion test was .76, while the KR-20 index for the visual scanning criterion test was .51.

In addition to the direct comparisons of the psychometric indexes of the two types of tests, additional comparisons were made which addressed the conclusions drawn from evaluations made of adaptive testing procedures using data gathered from live adaptive tests and computer simulations. In pairwise comparisons of the three adaptive tests, using both live and simulated data, comparable conclusions were arrived at by both data sources in five of six comparisons.

These results support the use of computer simulations as a practical tool for the exploratory evaluation of adaptive tests. However, the presence of a significant effect attributable to the administration of items in an adaptive as opposed to a conventional format, suggests that the results of such simulations should be regarded as tentative until empirical confirmation studies using live test administrations are performed.

These cautions are especially relevant as computer-based testing moves from laboratory to operational usage. The writings of Urry (1977b) and his colleagues (cf. McKillip, 1977) at the Civil Service Commission (1976), and scientists such as Ree (1977) and Cory (1975) in the Department of Defense, clearly show that the age of computer-based testing, probably using some adaptive testing format, is fast approaching. The time-savings of adaptive testing techniques, the resistance to test compromise, and the capacity to implement novel testing protocols, such as measures of perceptual and psychomotor abilities, all offer substantial advantages over conventional paper-and-pencil testing.

While the findings of the current research cannot be generalized beyond the specific tests and item types evaluated, the implications of dissimilar live and simulation results for a body of science in which computer simulations are increasing in popularity are evident. The present results have demonstrated significant differences between scores from live administrations of adaptive tests and scores from real-data based computer simulations of those same adaptive tests. Although small in magnitude, these effects nevertheless suggest the need for continued verification of simulation procedures. As Vale and Weiss (1975a, p. 5) stated, "A simulation study is valuable only to the extent that the underlying model accurately reflects data from live-testing studies."

REFERENCE NOTES

1. Maurelli, V. A. A comparison of Bayesian and maximum likelihood scoring in a simulated stradaptive test. (Unpublished master's thesis). San Antonio, Texas: St. Mary's University, May 1978:

2. Wood, R. Fully adaptive sequential testing: A Bayesian procedure for efficient ability measurement. Unpublished manuscript, 1972.

REFERENCES

Angoff, W. H., & Huddleston, E. M. The multi-level experiment: A study of a two-level test system (SR-58-121). Princeton, N.J.: Educational Testing Service, 1958.

Bayroff, A. G. Feasibility of a programmed testing machine. Research Study 64-3. Washington, D.C.: U.S. Army Behavioral Science Research Laboratory, November 1964.

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests (Research Note 188). Washington, D. C.: United States Army Behavioral Science Research Laboratory, June 1967.

Bejar, I. I. A comparison of conventional and computer-based adaptive achievement testing. In D.J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, October 1977.

Betz, N. E. Effects of immediate knowledge of results and adaptive testing on ability test performance. Applied Psychological Measurement, 1977, 1, 259-266.

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). Minneapolis: University of Minnesota, October 1973.

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis: University of Minnesota, October 1974.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minnesota, July 1975.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3). Minneapolis: University of Minnesota, June 1976.

Church, A. T., & Weiss, D. J. Interactive computer administration of a spatial reasoning test (Research Report 80-2). Minneapolis: University of Minnesota, April 1980.

Civil Service Commission. Computers and testing, steps toward the inevitable request (PB-261 694). Washington, D. C.: United States Department of Commerce, 1976.

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)

Cleary, T. A., Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187. (b)

Cliff, N., Cudeck, R., & McCormick, D. An empirical evaluation of implied orders as a basis for adaptive testing. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977.

Cory, C. H. Using computerized tests to add new dimensions to the measurement of abilities which are important for on-job performance: An exploratory study. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (Report 75-6). Washington, D. C.: United States Civil Service Commission, June 1975.

Cudeck, R. A., Cliff, N., & Kehoe, J. F. TAILOR: A FORTRAN procedure for interactable tailored testing. Educational and Psychological Measurement, 1977, 37, 767-769.

Cudeck, R. A., McCormick, D. J., & Cliff, N. Monte Carlo evaluation of implied orders as a basis for tailored testing. Applied Psychological Measurement, 1979, 3, 65-74.

DeWitt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement (Research Report 74-1). Minneapolis: University of Minnesota, January 1974.

Elwood, D. L., & Griffin, H. R. Individual intelligence testing without the examiner: reliability of an automated method. Consulting and Clinical Psychology, 1972, 38, 9-14.

English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 1977, 9, 158-161.

Ferguson, R. L. Computer-assisted criterion-referenced measurement (WP-41). Pittsburgh, Pa.: Pittsburgh University Learning Research and Development Center, 1969.

Fletcher, J. D. Computer applications in education and training: Status and trends (NPRDC TR-75-32). San Diego, Ca.: Navy Personnel Research and Development Center, April 1975.

Garrett, H. E. Statistics in psychology and education. New York: McKay, 1958.

Gialluca, K. A., & Weiss, D. J. Effects of immediate knowledge of results on achievement test performance and test dimensionality (Research Report 80-1). Minneapolis: University of Minnesota, January 1980.

Green, B. Comments on tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Greenwood, D. J., & Taylor, C. Adaptive testing in an older population. Journal of Psychology, 1965, 60, 193-198.

Guilford, J. P., & Fruchter, B. Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill, 1978.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hansen, D. H. An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

Hansen, D. H., Ross, S., & Harris, D. A. Flexilevel adaptive testing paradigm: Hierarchical concept structures (AFHRL-TR-77-35-II). Lowry Air Force Base, Colorado: Technical Training Division. Air Force Human Resources Laboratory, July 1977. (a)

Hansen, D. H., Ross, S., & Harris, D. A. Flexilevel adaptive testing paradigm: Validation in technical training (AFHRL-TR-77-35-I). Lowry Air Force Base, Colorado: Technical Training Division, Air Force Human Resouces Laboratory, July 1977. (b)

Hick, W. E. Information theory and intelligence tests. British Journal of Psychology, Statistical Section, 1951, 4, 157-164.

Holtzman, W. H. (Ed.). Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Hornke, J. F., & Sauter, M. P. A validity study of an adaptive test of reading comprehension. In D. J. Weiss (Ed.), Conference proceedings (Report 27-30). Minneapolis: University of Minnesota, June 1979.

Hunter, D. R. Development of an enlisted psychomotor/perceptual test battery (AFHRL-TR-75-60). Lackland Air Force Base, Texas: Personnel Research Division, Air Force Human Resources Laboratory, November, 1975.

Hunter, D. R. Research on computer-based perceptual testing. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977.

Hutt, M. L. A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

Ireland, C. M. An application of the Rasch one parameter logistic model to individual intelligence testing in a tailored testing environment. (Doctoral dissertation, University of Missouri, 1976). Dissertation Abstracts International, 1977, 37, 5766A.

Jensema, C. An application of latent-trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 29-48.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1977, 1, 111-120.

Johnson, M. F., & Weiss, D. J. Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.) Conference Proceedings (Report 27-30). Minneapolis: University of Minnesota, June 1979.

Kalisch, S. J. A model for computerized adaptive testing related to instructional situations. In D. J. Weiss (Ed.), Conference Proceedings (Report 27-30). Minneapolis: University of Minnesota, June 1979.

Kalisch, S. J. Computerized instructional adaptive testing model: Formulation and validation (AFHRL-TR-79-33). Lowry Air Force Base, Colorado: Technical Training Division, Air Force Human Resources Laboratory, February 1980.

Kingsbury, G. G., & Weiss, D. J. Relationships among achievement level estimates from three item characteristic curve scoring methods (Research Report 79-3). Minneapolis: University of Minnesota, April 1979.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis: University of Minnesota, July 1974.

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report 75-1). Minneapolis: University of Minnesota, February 1975.

Linn, R. L., Rock, D. A., & Cleary, T. A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F. M. The self-scoring flexilevel test (RB-70-43). Princeton, N.J.: Educational Testing Service, 1970.

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (a)

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-241. (b)

Lord, F. M. Robbins-Munro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (c)

Lord, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. (d)

Lord, F. M. A broad-range tailored test of verbal ability. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (Report 75-6). Washington, D. C.: United States Civil Service Commission, June 1975.

Lord, F. M. Applications of item response theory to practical testing problems. Hillside, N. J.: Earlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McBride, J. R. Adaptive testing research at Minnesota: Some properties of a Bayesian sequential adaptive mental testing strategy. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (Report 75-6). Washington, D. C.: United States Civil Service Commission, June 1975. (a)

McBride, J. R. Scoring adaptive tests. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, November 1975. (b)

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R. Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), Conference proceedings (Report 27-30). Minneapolis: University of Minnesota, June 1979.

McBride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement (Research Report 74-2). Minneapolis: University of Minnesota, June 1974.

McCormick, D. J., & Cliff, N. TAILOR-APL: An interactable computer program for individual tailored testing. Educational and Psychological Measurement, 1977, 37, 771-774.

McKillip, R. H. Implementation of tailored testing at the Civil Service Commission. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977.

Novick, M. R. Bayesian methods in psychological testing (RB 69-31). Princeton, N.J.: Educational Testing Service, 1969.

Owen, R.J. A Bayesian approach to tailored testing (RB-69-31). Princeton, N.J.: Educational Testing Service, 1969.

Pine, S. M. Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis: University of Minnesota, March 1977.

Prestwood, J. S. Knowledge of results and the proportion of positive feedback on tests of ability. Applied Psychological Measurement, 1979, 3, 155-160.

Ree, M. J. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examining Station. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977.

Ree, M. J. The effects of item calibration sample size and item pool size on adaptive testing. Applied Psychological Measurement, 1981, 5, 11-19.

Sanders, J. H., Valentine, L. D., Jr., & McGrevy, D. R., The development of equipment for psychomotor assessment (AFHRL-TR-71-40). Lackland Air Force Base, Texas: Personnel Research Division, Air Force Human Resources Laboratory, July 1971.

Sapinkopf, R. C. A computer adaptive testing approach to the measurement of personality variables. (Doctoral dissertation, University of Maryland, 1977), Dissertation Abstracts International, 1978, 38, 4993B.

Schmidt, F. L., Urry, V. W., & Gugel, J. F. Computer assisted tailored test; examinee reactions and evaluations. Educational and Psychological Measurement, 1978, 38, 265-273.

Sympson, J. B., Weiss, D. J., & Ree, M. J. Predictive validity of conventional and adaptive tests in Air Force training (AFHRL-TR-81-in press). Brooks Air Force Base, Texas: Manpower and Personnel Division, Air Force Human Resources Laboratory, 1981.

Thompson, J. G., & Weiss, D. J. Criterion-related validity of adaptive testing strategies (Research Report 80-3). Minneapolis: University of Minnesota, June 1980.

Urry, V. W. A monte carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970), Dissertation Abstracts International, 1971, 31, 6319B.

Urry, V. W. A multivariate model sampling procedure and a method of multidimensional tailored testing. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, July 1977. (a)

Urry, V. W. Tailored testing: A successful application of latent-trait theory. Journal of Educational Measurement, 1977, 14, 181-196. (b)

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. Methods for linking item parameters (AFHRL-TR-81-10). Brooks Air Force Base, Texas: Manpower and Personnel Division, Air Force Human Resources Laboratory, August 1981.

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, December 1975. (a)

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, October 1975. (b)

Votaw, D. F. Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics, 1948, 19, 447-473.

Waters, B. K. An empirical investigation of Weiss' stradaptive testing model. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (Report 75-6). Washington, D. C.: United States Civil Service Commission, June 1975. (a)

Waters, B. K. Empirical investigation of the stradaptive testing model for the measurement of human ability (AFHRL-TR-75-27). Williams Air Force Base, Arizona: Flying Training Division, Air Force Human Resources Laboratory, October 1975. (b)

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

Waters, C. W., & Bayroff, A. G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, September 1973.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, December 1974.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, February 1973.

Wood, R. The efficacy of tailored testing. Educational Research, 1969, 11, 219-222.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

APPENDIX A: SAMPLE WORD KNOWLEDGE ITEM

WHICH OF THE FOLLOWING MEANS NEARLY THE SAME AS: FLOG

A. BEAT
B. WAVE
C. WRITHE
D. RESOUND
E. YIELD

___*___

*Examinee enters letter, followed by Carriage Return on the computer keyboard.

SAMPLE B: SAMPLE VISUAL SCANNING ITEM

<u>DISPLAY 1</u> (5 Seconds)

IN THE FOLLOWING DISPLAY COUNT THE NUMBER OF

TIMES THE TARGET NUMBER APPEARS

TARGET NUMBER = <u>4</u>

<u>DISPLAY 2</u> (7 Seconds)

---

1 5 8 4 9 4 6 8 2 4 2 7 9 3 5 7 2 7 3 4

5 8 2 4 2 6 8 9 3 6 7 4 4 7 9 2 4 7 8 2

<u>DISPLAY 3</u> (No Time Limit)

ENTER THE NUMBER OF TIMES THE TARGET NUMBER APPEARED

<u>*</u>

---

*Examinee enters response (1 or 2 digits), followed by Carriage Return on the computer keyboard.